 DOI: https://doi.org/10.12349/iser.v6i5.5700

Research Progress on the Autonomous Deployment of Lightweight AI Models Based on Edge Computing

Yuhan Dong[1] Chaoran Xing[1] Yixian Mei[1] Junbo Wang[2] Baojia Xu[1]

1. Chengxian College, Southeast University, Nanjing, Jiangsu, 210023, China
2. Zhongbei College of Nanjing Normal University, Nanjing, Jiangsu, 210088, China

Abstract

With the wide application of edge computing in fields such as autonomous driving and industrial Internet of Things, the deployment of lightweight AI models has become the key to achieving low-latency and high-efficiency autonomous intelligence. This paper systematically reviews the technological progress of model compression, hardware co-optimization and autonomous deployment tools: At the algorithmic level, pruning, quantization and knowledge distillation techniques can significantly compress the model size, and dynamic inference and lightweight Transformer architecture improve the inference speed at the edge. At the hardware collaboration level, the dedicated NPU and compilation toolchain significantly improve the energy efficiency ratio. Typical applications have verified the effectiveness of lightweight technology in scenarios such as industrial defect detection and wearable medical care. However, the balance between precision and efficiency, hardware fragmentation and adaptability to dynamic environments remain the core challenges. In the future, it is necessary to integrate directions such as automated compression tools, edge-cloud collaborative inference, and in-memory computing chips to promote the practicality and standardization of the edge intelligent ecosystem. This article provides technical references and practical guidelines for building an efficient and robust autonomous edge AI system.

Keywords

Edge Computing AI model Autonomous deployment

基于边缘计算的轻量化 AI 模型自主化部署的研究进展

董禹含[1] 邢超然[1] 梅益仙[1] 王俊博[2] 徐宝佳[1]

1. 东南大学成贤学院，中国 · 江苏 南京 210088
2. 南京师范大学中北学院，中国 · 江苏 南京 210023

摘　要

随着边缘计算在自动驾驶、工业物联网等领域的广泛应用，轻量化AI模型部署成为实现低延迟、高能效自主智能的关键。本文系统综述模型压缩、硬件协同优化及自主化部署工具的技术进展：在算法层面，剪枝、量化与知识蒸馏技术可将模型尺寸大大压缩，动态推理与轻量Transformer架构使边缘端推理速度提升；在硬件协同层面，专用NPU与编译工具链显著提升能效比。典型应用验证了轻量化技术在工业缺陷检测、可穿戴医疗等场景的实效性。然而，精度-效率平衡、硬件碎片化及动态环境适应性仍是核心挑战。未来需融合自动化压缩工具、边缘-云协同推理及存算一体芯片等方向，推动边缘智能生态的实用化与标准化。本文为构建高效、鲁棒的自主化边缘AI系统提供技术参考与实践指南。

关键词

边缘计算；AI模型；自主化部署

1 引言

随着工业物联网与自动驾驶等实时智能应用的发展，边缘计算凭借低延迟、隐私保护等优势成为关键支撑技术。然而，边缘设备受限于算力、内存与能耗，难以直接部署复杂 AI 模型 [1]。轻量化 AI 技术通过模型压缩、动态架构与硬件协同实现高效边缘部署。此研究可以从单一算法优化转向“算法 - 硬件 - 系统”全栈协同，显著提升工业检测、可穿戴医疗等场景的实用性。然而，精度 - 效率权衡、硬件碎片化适配与动态环境适应仍是核心挑战。本文系统性综述轻量化模型部署的技术进展，解析模型压缩、自主架构与工具链优化的关键方法，结合典型应用与性能数据验证技术实效性，并探讨自动化轻量化、边缘 - 云协同等未来方向，为构建鲁棒、高效的边缘智能系统提供参考。

【作者简介】董禹含（2006–），女，中国江苏徐州人，本科，从事计算机科学与技术研究。

2 轻量化模型关键技术

2.1 算法层压缩

为实现边缘设备上 AI 模型的高效部署，算法层压缩技术通过去除冗余参数、降低计算精度与迁移知识，显著缩减

模型规模。本节从剪枝、量化与知识蒸馏三方面综述技术进展。

2.1.1 剪枝：结构化与非结构化策略的自主权衡

剪枝技术通过移除冗余连接或神经元降低模型复杂度，其核心方法包括：

（1）结构化剪枝：结构化剪枝通过删除神经网络中完整的结构单元，在保持模型结构规则性的同时降低参数量与计算量，提升推理效率。有相关研究提出基于深度神经网络二阶信息的结构化剪枝算法，实验结果表明，该算法在网络参数量和每秒浮点运算次数分别减少 29.9% 和 34.6% 的情况下，在 ResNet110 网络上的分类准确率提升了 0.74%，剪枝效果优于 PF、LCCL 等经典剪枝算法 [2]。

（2）非结构化剪枝：非结构化剪枝则不遵循特定的结构模式，直接对模型中的单个参数进行裁剪。这种方式能够在理论上达到更高的压缩率，但非结构化剪枝后模型的稀疏性在现有规则架构中难以高效实现，因为硬件通常更适合处理密集的数据结构 [2]。

2.1.2 量化：低比特整型推理的端侧适配

量化通过将 32 位浮点权重映射至低精度整数（如 8/4 位），降低内存与计算开销。关键进展包括：

（1）整数量化实践：整数量化实践指将 32 位浮点权重转换为 8/4 位等低精度整数的具体操作，通过设计映射与反量化算法，在保障模型精度前提下，实现存储与计算效率显著提升。

（2）自主量化策略：自主量化策略是基于算法自动探索最优量化方案，动态调整量化参数与粒度，平衡模型压缩、推理速度与精度，减少人工调优成本，增强量化过程自适应性。

2.1.3 知识蒸馏：自动化师生架构生成

知识蒸馏通过迁移教师模型知识训练轻量学生模型，其自主化演进方向包括：

（1）自动化框架设计：自动化设计框架借助算法与工具，替代人工完成设计。通过规则与优化算法，提升效率与精度，广泛应用于多领域，推动设计智能化。有相关研究提出了一种基于“教导主任 - 教师 - 学生”模式的统一的模型压缩技术，该方法能够动态结合蒸馏学习和结构化稀疏裁剪，通过学习可蒸馏性和可稀疏性，自适应地调节联合训练机制，可作为自动化框架设计方面的参考。[3]

（2）联邦蒸馏：联邦蒸馏融合联邦学习与知识蒸馏，解决联邦学习中通信成本高、架构受限问题。它突破传统参数共享模式，实现客户端与服务器间灵活的知识传递，降低训练开销。通过传输预测信息或压缩伪样本等方式，在不共享数据的前提下提升模型性能。

2.2 自主化架构设计

为实现边缘设备上 AI 模型对动态环境的自主适应，本节从动态推理机制与轻量 Transformer 架构两个层面重构网络设计范式，使模型能够根据输入复杂度与资源状态实时调整计算负载。

2.2.1 动态推理：实时计算路径选择

动态推理通过构建多分支网络，允许模型在不同输入条件下自主选择计算路径。其核心方法包括：

早退机制（Early Exiting）：早退机制在神经网络浅层设置多个推理 “出口”，模型在中间层计算过程中实时评估预测结果的置信度。当某中间层输出的预测置信度满足预设阈值时，可提前终止后续深层计算，直接输出结果，从而减少不必要的计算开销，提升推理效率。

条件计算（Conditional Computation）：条件计算基于输入数据的特征分布，动态激活神经网络中部分神经元或分支。相较于传统模型对所有输入均执行完整计算流程，该方法可根据输入特性灵活启用相关计算模块，跳过冗余运算，在保障推理精度的同时，显著降低计算量，优化推理效率。

2.2.2 轻量 Transformer：自适应注意力优化

传统 Transformer 因自注意力机制的高计算复杂度难以部署至边缘设备。轻量化改进聚焦于注意力稀疏化与跨模态效率平衡，通过动态调整计算负载与模型结构适配边缘资源约束。

（1）自适应注意力头剪枝：自适应注意力头剪枝依据输入特征重要性，动态关闭冗余注意力头。通过量化各头贡献度，设定动态阈值跳过非关键计算，在减少冗余开销的同时，通过阈值灵活平衡精度与效率。

（2）卷积 - 注意力混合架构：卷积 - 注意力混合架构融合卷积局部特征提取与自注意力全局建模优势，设计轻量模块。浅层卷积捕捉细节，深层稀疏注意力建模全局关系，复用参数减少冗余，以低计算量实现多任务适配。

（3）序列长度压缩：序列长度压缩针对序列数据，采用时间维度下采样或特征池化缩短输入序列。借助卷积 / 池化层、局部注意力计算等技术，降低内存占用，减少计算量，有效提升模型在语音、文本等场景的实时性。

3 应用场景与挑战

3.1 典型应用场景

3.1.1 工业物联网：实时缺陷检测

在智能制造领域，轻量化 AI 模型可部署至产线边缘设备（如工业相机、PLC 控制器），实现高精度实时质检。

3.1.2 智能家居：本地化语音交互

为保护用户隐私并降低云端依赖，轻量化语音识别模型可部署至智能音箱、家电控制器等设备。

3.1.3 智慧医疗：可穿戴健康监测

轻量化时序模型在可穿戴设备中实现实时生理信号分析。

3.2 核心挑战

3.2.1 精度与效率的强耦合性

轻量化技术往往以牺牲模型精度为代价换取效率提升。此外，动态剪枝技术虽能自主调整压缩率，但在资源波动剧烈的场景（如移动设备电量不足时），频繁的模型切换可能引发精度不稳定。

3.2.2 硬件生态碎片化

边缘设备硬件架构高度异构。由于各架构在计算能力、功耗特性和编程接口上缺乏统一标准，跨平台适配需深度优化，导致开发成本显著提升，严重阻碍边缘计算应用的规模化部署。

3.2.3 动态环境适应性不足

现有动态推理机制依赖预设阈值，难以适应复杂多变的运行环境。例如在智能交通场景中，数据流量的时变性会使固定阈值策略的推理效果大幅降低。同时，边缘 - 云协同受网络条件制约明显，网络波动时数据传输延迟激增，导致协同推理失败率提高。

4 未来方向

4.1 自动化轻量化技术

4.1.1 端到端自动化压缩流水线

当前模型压缩技术依赖人工分阶段调参，效率较低。未来研究将聚焦于构建端到端自动化压缩流水线，通过算法集成与优化，实现压缩流程的自动化，降低人力成本，提升模型压缩效率与适配性，推动边缘计算模型的快速部署。

4.1.2 硬件感知的神经架构搜索（NAS）

传统神经架构搜索（NAS）方法在模型设计时，往往忽视硬件特性约束，导致搜索得到的模型虽理论性能优异，但在实际边缘设备上难以高效运行。硬件感知的 NAS 需将设备算力、内存、功耗等硬件参数深度融入搜索过程，构建软硬件协同优化机制，从而生成适配边缘设备的高效模型架构 [4]。

4.1.3 轻量化联邦学习

联邦学习旨在保护数据隐私的同时，实现多设备协同训练，但在边缘场景下，其通信开销与模型轻量化需求间存在突出矛盾。为保证训练效率与隐私安全，需探索新型轻量化算法，如优化参数传输策略、设计低维模型结构，平衡通信成本与模型性能，推动联邦学习在边缘计算中的实用化进程 [5]。

4.2 边缘 – 云协同推理

4.2.1 动态模型分割与卸载

边缘设备算力有限，面对复杂推理任务时，需与云端协同处理。传统固定分割策略难以适应动态负载变化，未来需研究动态模型分割算法，依据设备实时算力、网络状态智能划分计算任务，实现模型的灵活卸载与高效协同，在降低边缘计算压力的同时保障推理时效性 [6]。

4.2.2 增量学习与在线适应

边缘设备需利用本地数据持续优化模型以适应环境变化，但直接更新易引发隐私泄露风险。需探索安全增量学习框架，通过加密技术与隐私保护机制，在保障数据隐私前提下，实现模型参数的安全更新与在线自适应，提升模型在复杂边缘场景下的泛化能力 [7]。

4.2.3 数据协同与联邦推理

多边缘节点数据协同推理可有效提升全局模型鲁棒性。然而，节点间数据异构、通信延迟等问题制约协同效率。需设计高效联邦推理协议，平衡数据隐私保护与信息共享，优化节点间通信机制，实现多源数据融合与模型协同更新，增强边缘计算系统整体性能与可靠性。

4.3 新型计算范式融合

4.3.1 光子计算加速

光子计算基于光信号并行传输与处理特性，相比电子计算具备天然高速、低功耗优势。其数据处理的高并行性与低延迟特点，契合边缘端对实时性和能效的严苛要求，尤其适用于视频处理、高速数据采集等场景，可提升边缘计算系统的响应速度与处理能力。

4.3.2 类脑计算与脉冲神经网络

类脑计算借鉴生物神经系统结构，以脉冲神经网络模拟神经元电脉冲活动，采用事件驱动机制，仅在关键事件发生时激活计算，具备极低功耗与强大稀疏数据处理能力。该技术能有效降低边缘设备能耗，在物联网感知、实时监测等场景展现出独特的应用潜力。

5 结论

边缘计算驱动的轻量化 AI 模型自主化部署技术，通过算法层压缩、动态推理架构优化以及硬件协同设计，显著降低了模型复杂度与计算开销，在工业检测、可穿戴设备等场景中实现了推理效率与能效比的同步提升，同时保持较高模型精度。然而，该领域仍面临精度与效率的动态平衡难题、异构硬件适配的高复杂度挑战以及动态环境下模型鲁棒性不足等瓶颈。未来研究需围绕三方面展开：一是构建自动化轻量化技术体系，深度融合硬件感知的模型压缩与跨平台适配工具；二是推动边缘 - 云协同推理框架创新，强化动态资源调度与隐私保护机制；三是探索存算一体、光子计算等新型架构，突破传统算力与能效限制。这些方向的发展将为边缘智能在工业物联网、智慧医疗等领域的规模化应用提供理论支撑与技术保障，促进高效、安全、自适应的边缘 AI 生态系统构建。

参考文献

[1] 李辉,李秀华,熊庆宇,文俊浩,程路熙. "边缘计算助力工业互联网:架构、应用与挑战."《计算机科学》1(2021).

[2] 季繁繁,杨鑫,袁晓彤. "基于深度神经网络二阶信息的结构化剪枝算法."《计算机工程》2(2021).

[3] Xizi Chen,Jingyang Zhu,Jingbo Jiang,C. Tsui. “Tight Compression: Compressing CNN Model Tightly Through Unstructured Pruning and Simulated Annealing Based Permutation.” 2020 57th ACM/IEEE Design Automation Conference (DAC)(2020).

[4] 王鑫,姚洋,蒋昱航,关超宇,朱文武. "硬件感知的神经架构搜索."《中国科学:信息科学》5(2023).

[5] 王周生,杨庚,戴华. "基于梯度选择的轻量化差分隐私保护联邦学习." 计算机科学 1(2024).

[6] 刘耕,唐小勇,游正鹏,赵立君,陈庆勇. "联邦学习在5G云边协同场景中的原理和应用综述." 通讯世界 7(2020).

[7] 任建吉,王鹏然. "动态物联网环境下的联盟学习计算卸载优化." 计算机工程与应用 16(2019).